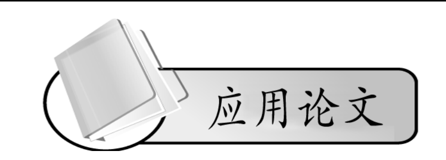



# 基于CSpace的科技信息可配置化自动监测功能设计与实现*

王思丽[1,2] 刘 巍[1] 祝忠明[1] 吴志强[1] 王金平[1]
[1](中国科学院兰州文献情报中心 兰州 730000)
[2](中国科学院大学 北京 100049)

**摘要:**【目的】实现对多源异构科技信息的长期监测、自动采集发布与存储管理, 以满足专题领域科技研究的需求。【方法】结合 CSpace 的应用扩展需求, 设计开发了基于 CSpace 的可配置化的科技信息自动监测功能, 着重研究和解决了多源异构科技信息采集内容规则的可配置化实现、与 CSpace 交互的自动采集发布接口的可配置化实现等关键技术问题, 并以海洋科技信息的自动监测采集为例进行应用研究。【结果】能够实现对多源异构科技信息的自动监测采集, 为科技平台建设提供良好支持。【局限】采集内容规则配置过程比较复杂; 不支持对一些需要登录的复杂站点的自动监测。【结论】该功能方法较大程度上扩展了 CSpace 的数据采集集成功能, 且具有一定的通用性、可配置性与松耦合性, 可应用于多个科技信息监测领域。

**关键词:** CSpace 机构知识库 科技信息 自动监测 信息采集
**分类号:** TP274.2
**DOI:** 10.11925/infotech.2096-3467.2017.0783



## 1 引 言

在现代开放信息环境下, 网络上的科技信息资源由于其时效性强, 覆盖范围广, 且一定程度上具有较大的可信度(尤其是权威机构发布的), 已成为情报研究人员关注的重点, 及时发现、分析、管理和利用这些科技信息资源, 对于获得最新的情报信息, 制定合理的科技战略规划, 进行相关情报决策研究十分必要。本文出于项目建设和 CSpace 应用的双向需求, 基于CSpace进行功能扩展开发, 使其能够实现对网络中开放性科技信息的自动监测和存储分析管理。

CSpace 是中国科学院机构知识库(Institutional Repository)建设平台，目前主要应用于对机构产出的各种科研成果进行存缴管理、集成共享和长期保存。但 CSpace 除了具备支持常规机构知识库建设管理的全系列功能模块外, 它的动态元数据框架和知识对象类型化模板化机制, 非常有利于支持多类型信息资源的采集与集成, 能够适应多场景多领域下的专题性数字知识库的建设[1-3]。目前 CSpace 对网络科技信息资源采集集成的支持主要有以下两种方式:

(1) 可通过批量导入接口直接导入已采集的信息。但导入之前需要人工下载导入模板, 按模板要求严格进行数据预处理, 实现已有数据信息和CSpace元数据字段的映射。

(2) 提供可配置化的 OAI-PMH 接口对支持标准 OAI-PMH 协议的信息源进行定时收割聚合; 提供可

通讯作者: 王思丽, ORCID: 0000-0002-2126-3462, E-mail: wangsl@llas.ac.cn。

*本文系中国科学院西部之光青年学者A类项目"基于学术大数据的专题化信息自动采集与组织技术研究"(项目编号: Y6AX021001)和中国科学院西部之光青年学者 B 类项目"科研过程中非文本资源的语义化组织技术研究"(项目编号: Y6AX031001)的研究成果之一。

视化的采集服务界面对拥有 WOS Web Service 服务权限的机构用户提供数据采集功能。收割和采集的数据会根据配置自动提交与关联映射到 CSpace 知识库相应的研究专题和知识内容类型中。

这两种方式在对科技情报源进行监测采集时都存在不足。首先，需要借助于大量的人工处理，动态交互性和时效性都不够强。而科技情报源的监测对信息的精准性和时效性都有较高的要求，且大多不提供批量下载功能，原始数据的实时采集也是一个问题，没有批量下载就难以批量上载。其次，科技情报源与文献型数据库不同，来源站点一般是综合性门户性网站，大多都不支持不提供标准的数据互操作协议和接口，因而并不具有普适性。但同时，科技情报源的监测在实现上也具有一定优势:

(1) 由于其信息来源一般是重要国家的重要机构性网站，信息来源是明确的、有限的，因此信息源可以预先遴选和配置。

(2) 信息源的内容结构虽然复杂多样，层次不一，但基本都是一个站点多个栏目下的概览页(列表页)–详细信息页(正文内容页)的构造模式，有利于研究实现通用的可配置化的基于多来源站点多栏目内容信息的精准定位采集。

本文在对网络科技信息监测的相关研究现状进行调研梳理的基础上，结合 CSpace 自身的架构模式和科技情报监测的建设需求，设计开发了基于 CSpace 的可配置化的科技信息自动监控功能，着重研究和解决了多来源多栏目科技情报源采集内容规则的可配置化实现，与 CSpace 交互的自动采集发布接口的可配置化实现等关键技术问题，并以海洋科技信息的自动监测采集为例进行应用实践，最终实现了基于 CSpace 对多源异构科技信息的长期监测、自动采集发布与存储管理。并且该功能方法具有通用性和可配置性，也可以用于相关专题领域科技资源的自动监测与采集建设。

## 2　研究综述

网络科技信息监测与普适性的搜索引擎系统所关注的采集目标有所不同，属于主题信息采集的范畴。自 2000 年以来，国内外主题信息采集技术愈来愈成熟，逐渐得到广泛研究和深入应用，所涉及到的相关技术一般包含采集规则/算法/模型的构建、主题内容信息的自动识别和抽取、网页文本的自动聚类与分类技术等。从主要的技术实现方式上大致可分为 5 类:

(1) 基于 URL 规则的方法

该方法主要基于同一个来源站点创建的动态网页其内容一般应属于同一个主题且其 URL 往往非常相似这一规律，通过各种算法和模型实现对这一规律的量化、补充计算，以区分主题无关的 URL 和主题相关的 URL。如叶勤勇[4]提出 UFBC 学习算法，基于开源搜索引擎 Nutch 和利用正则表达式进行算法实现; 蒋付彬[5]提出的基于决策树的 URL 分类器算法，利用 4 个主要 HTML 标签内容与用户定义主题的相似度构建决策树实现 URL 分类; 杨镒铭[6]提出基于模式树的 UPCA 分类算法，通过训练提取特定类型的网页链接特征，构建模式树和生成模式规则，形成主题相关的 URL 模式库。

(2) 基于模板匹配的方法

该方法主要基于同一网站其内容页面都基于相同的模板这一规律，首先创建和识别模板，然后基于模板进行主题信息内容匹配抽取。如 Bar-Yossef 等[7]将网页的头部、侧部等导航栏、底部版权声明、广告等网页中公有的重复出现的信息视为噪音信息并定制为模板，基于网页 DOM 树和模板对待处理网页进行匹配删除，最后剩下的为主题相关的信息。

(3) 基于机器学习的方法

该方法一般需要通过大量的样本积累和训练，或者由人工预先标注好一定数量的样本实例，交给机器程序去聚类、归纳学习，生成网页分类器(算法和规则)，利用分类器对网页信息进行模式处理。如 Mitra 等[8]利用预定义的标签集合对 DOM 树节点进行训练，生成分类器; 王浩[9]提出了将采样技术和半监督学习相结合的方法，对传统的 SMOTE 文本分类算法进行改进以实现网络敏感信息的识别; Pavlinek 等[10]提出了基于主题模型表示的半监督式文本分类方法，该方法包括一个基于自训练的半监督文本分类算法和模型，用于识别和确定新文本内容的参数设置。

(4) 基于启发式规则的方法

该方法主要结合网页的内容结构特征和视觉特征，采用相关启发式算法如神经网络算法、贪心算法等构建启发式规则集合，将网页划分为多个可视化块的相关集合以实现内容信息提取。如李剑[11]基于 BP 神经网络算法改进 DOM 树结构，按内容相关性将网

页划分为多个子模块进行信息内容过滤提取; 李伟男等[12]基于模拟退火算法训练二阶隐马尔科夫参数, 改进经典的 VIPS 网页分块算法[13], 以实现网页主题信息抽取; 谢方立[14]提出了基于 DOM 节点类型标注的 NTA 主题信息抽取算法。

(5) 其他综合性技术方法

将上述几种技术方法同数学、计算机、图书情报等领域的各种方法如向量空间模型 VSM、泊松分布模型、贝叶斯分类算法、模糊数学方法、大数据/云计算技术、知识发现技术等选择性地、有侧重点地结合起来, 以解决信息监测中出现的各种问题。如欧健文等[15]对基于模板的匹配方法进行改进, 提出基于机器学习的线性回归算法生成模板, 通过检测链接之间的关系和识别锚文本的特征建立页面模板及提取规则, 实现了对网页信息主体的识别和提取; 马费成等[16]采用模糊数学的理论方法, 构建了网页生命阶段识别指标和模糊识别模型, 以实现对网络信息生命状态的定位, 从而判断出网页采集更新的最佳周期和策略; 林文辉[17]对基于 Hadoop 和云计算分析的网络数据采集和处理的关键技术进行研究。

上述这些技术方法在实现应用时各有侧重点, 并不是一个技术由高到底被取代的过程, 而是综合利用、相辅相成的。在应用研究时, 一般和优秀的开源网络搜索引擎系统和网络爬虫框架如 Lucene、Nutch、Heritrix、Crawler4j、Scrapy 等结合起来进行开发实现。如谭宗颖等[18]基于网络爬虫技术和文本聚类技术构建了科技发展前沿信息监测与分析平台; 刘海波[19]基于 Ajax 和 Web Service 技术实现了网站多栏目多频道的信息监测和实时入库; 张智雄等[20]构建了一种支持按需申请、定制服务的科技战略监测服务云平台, 通过将网络自由文本转化为结构化的可计算的知识单元, 实现对科技领域的态势监测; 谢靖等[21]以开源爬虫 Crawler4j 为基本框架, 实现了面向网络科技监测的分布式定向资源精确采集; 王思丽等[22]也在前期对开放资源的元数据自动采集策略方法进行实验研究。

上述技术和应用方法为本文的研究提供了思路。

## 3 关键功能设计与实现

### 3.1 整体功能结构

整体功能结构主要包含数据准备、采集内容规则的可配置化、自动采集发布接口的可配置化实现三层关键工作流程, 其框架如图 1 所示(自下而上)。

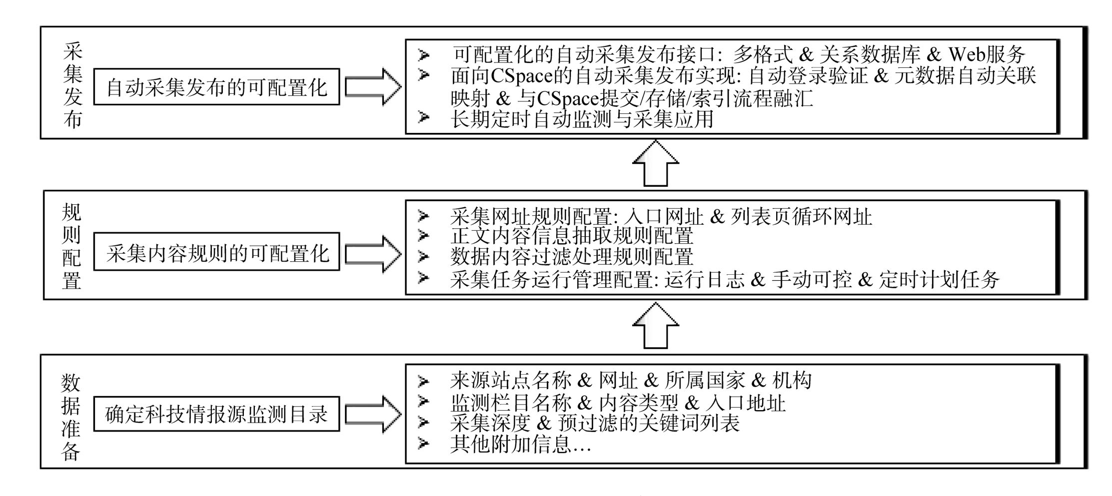



图 1　功能框架

### 3.2 数据准备

科技情报源监测目录的确定与准备。可由相关人员先行遴选和梳理提供, 主要包含监测所需的来源站点名称、站点网址、栏目名称、栏目入口地址、栏目内容类型、采集深度、预过滤的关键词列表等, 除了上述必提供项, 也可根据具体建设需求, 附加一些需要在采集配置时缺省批量写入的元数据项, 如来源机构名称、所属国家等。一个来源站点可允许提供并监测配置多个栏目及内容类型。

### 3.3 采集内容规则的可配置化实现

根据已遴选和确定的科技情报源监测目录, 对其采集内容规则进行可配置化实现, 包含采集网址规则

(采集入口网址规则、列表页循环网址规则)、正文内容信息抽取规则、数据内容过滤处理规则、采集任务运行管理配置等。其核心策略流程如图 2 所示。

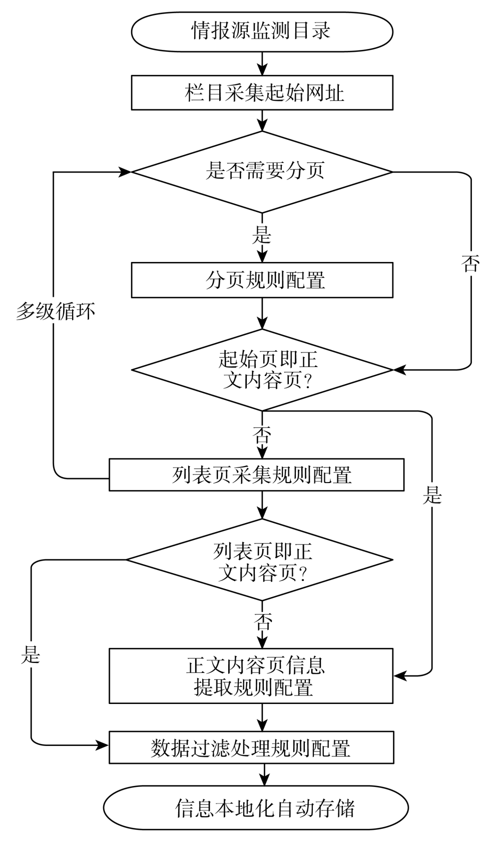



图 2　采集内容规则的配置流程

(1) 基于 URL 相似规律的分页规则配置

采用基于 URL 规则的方法，寻找分页 URL 的相似性规律。研究发现，分页 URL 地址格式一般可分为固定前缀部分和可变参数部分。其中可变参数部分通常遵循 4 类规则:

①等差数列规则。如果首页参数为 $p1$，公差为 $d$，那么第 $n$ 页的参数表达式为 $pn=p1+(n-1)\times d$。常见的分页参数公差为 1 或 10。首页参数为 0 或 1，项数可根据网站栏目提供的总页数进行确定。

②等比数列规则。如果首页参数为 $p1$，公比为 $q$，那么第 $n$ 页的参数表达式为 $pn=p1\times q^{n-1}$。

③A-Z 或 a-z 的字母变化规则。

④基于时间组合格式的变化规则。一般是年(yyyy)、月(MM)、日(dd)、时(HH)、分(mm)、秒(ss)的各种形式再加以分隔符的组合，如以下划线“_”、中杠“-”、反斜杠“/”等。可以通过格式化日期函数实现各种组合。此外，通常若首页的日期参数格式是 D1，则第 $n$ 页的日期参数格式可能比首页日期向前推迟几天。此认知基于大部分来源站点栏目信息数据最新的信息排列在最前页这一规律。

本文主要对上述 4 类规则进行配置化实现，在用户输入完整采集网址后，提示用户保留固定前缀部分，对于可变参数部分用(*)代替，并提供 4 种规则选项给用户进行选择和参数配置，系统最终会生成一个分页规则逻辑表达式进行规则存储，并将逻辑表达式可能监测到的 URL 地址列表提供给用户进行预览，以修正配置规则和参数，确保将监测的基本范围控制在某个栏目内。

(2) 基于多级循环和模板脚本匹配的列表页采集配置

通过对科技情报源的样本数据分析，基于迭代循环和逐层链接访问的方法实现对站点列表页的多级循环设计，设计流程如图 3 所示。

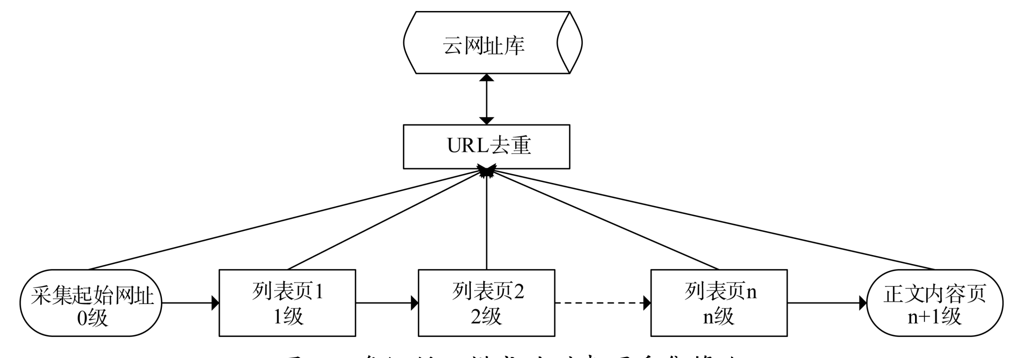



图 3　多级循环模式的列表页采集策略

其核心思想是将采集起始网址(包括其分页地址)视为 0 级网址，认为 0 级网址内应至少包含一个列表页 1，而 1 级列表页网址范围内的每一个链接可能同样包含了一个列表页 2，2 级列表页内的网址可能指向列表页 3，依次迭代循环下去，最终必定指向正文内容页。这个迭代的次数就是采集的深度，理论上可以

实现无限级迭代，但在实际应用中，受采集效率及外在网络的影响，一般设置迭代次数不超过 4 次，常见的为 2 级采集。对于确实需要深度迭代的可以切分为多个任务依次执行。同时，构建云网址库，初始化时云网址库为空，采集任务执行时首先将 0 级网址的第一个监测网址加入网址库，然后每次将循环采集到的列表网址与云网址库进行比对以实现 URL 去重和增量采集。

在对列表页内容进行采集规则配置时，主要采用改进的模板脚本匹配的方法，基于列表页中的数据列表一般都包含在一个固定的标签区域块中，且数据列表中的每一条概览性信息一般都具有大致相同的内容结构这一规律进行实现。在配置界面中，着重实现 5 种方式供用户进行规则选择:

①从页面中自动分析以得到列表页中的链接。主要基于网址链接都存储在 a 标签的 href 属性中这一内容结构特征。

②手动填写链接地址规则。用户需要基于 HTML 标签和内容结构自行建立模板脚本规则，并通过设置参数指定下一层级访问的实际链接地址。

③采用 XPath 方式获取地址链接。用户可以直接填写 XPath 表达式，也可以通过点击系统内嵌的微型 XPath 浏览器，辅助测试构造 XPath。

④指定区域提取网址。主要基于字符串截取的方法，根据用户指定的开始区域和结束区域标签进行网址的提取，指定的区域必须是页面中唯一的。

⑤指定结果网址集过滤的关键词列表。可以配置必须包含的关键词列表和不得包含的关键词列表，以实现对采集网址自身的过滤控制。

以上 5 种方式前三种属于必选一种的单选规则，后两种属于复选规则，可以和前三种进行规则组配，以实现多级规则控制。此外，在 http 请求方式中，实现了基于 get 和 post 两种方式进行请求。一般其中若请求方式选择为 post，需要配置 post 发送的数据模板，post 数据一般也可包含固定部分和可变部分。可变部分又分为随机数值和分页参数数值，其中实现方式与分页规则的配置实现方式类同。

(3) 基于模板创建和标签定位的正文内容页信息提取配置

同一来源站点同一栏目内容类型的正文内容页一般涵盖大致相同的几种信息元素(元数据)，且相同元素的标签位置基本是固定不变的。可根据该特征预先创建信息模板，如常见的正文信息元素有标题、作者、发布日期、正文内容等，然后为每一个元素指定信息提取方式。本文着重设计并实现了 4 种信息提取配置方式:

①基于开始和结束字符串的前后字符串截取的信息提取方式。其中开始字符串必须是页面中唯一的，结束字符串必须是继开始字符串之后页面中唯一的。

②基于正则表达式的模式匹配的信息提取方式。正则表达式的组件可以是单个的字符、字符集合、字符范围、字符间的选择或者所有这些组件的任意组合。它作为一个模板，将待提取的信息字符构造模式与所搜索的正文内容字符串进行匹配。

③基于 XPath 的可视化信息提取方式。与列表页的获取地址链接时的 XPath 方式类似。

④基于标签内容组合的信息提取方式。该方法主要是指将通过上述三种方法得到的标签内容经过一定的数据处理后，重新组合为新的元数据内容。可以选择一个或多个已获取的标签内容，根据需要自由设定分隔符进行组合和数据内容处理。

同时，本文实现了多种数据内容处理方式供信息提取后进行按需调用，如 html 标签过滤方式，对不需要的 html 标签内容进行可选过滤，包括 script 脚本、frame 框架、首尾空白字符等。如自动提取关键词等，通过采用分词组件和算法实现自动分词后，设定分隔符链接前 1 至 5 个高频词作为关键词。其他的如字符截取、字符内容替换、基于正则表达式的内容替换、编码自动识别转换等。此外，还可以配置关键词列表，对正文内容实现过滤删除，与上文采集网址的关键词列表配置类似。具体采集时，当判断出正文内容不符合关键词列表要求时，无论其他信息是否已提取，该条信息都将会被自动跳过。

(4) 基于主题任务树和可定制任务计划的采集任务运行配置

主要采用带复选框的可收缩的主题树方式进行采集任务的运行管理。在创建采集任务初期，支持根据 CSpace 专题研究需要建立不同的主题层级，形成主题任务树。然后针对不同更新周期的信息来源站点主要实现了两种采集任务运行管理方式：其一是人工实时监控采集；其二是定时自动采集，通过预先制定采集任务计划，由机器自动执行采集任务。在方法一中，可根据实际需要选择一个或多个任务到任务队列池进行采集，可选择单步执行采集，如只采集栏目的网址或内容，也可以选择执行全部任务流程，从采集网址到

内容到自动提交发布一次完成。同时，可以在日志模块中，查看实时采集的日志输出信息。方法二中，实现了计划任务管理器，支持预先从主题任务树中选择一个或多个任务，灵活设置采集任务的执行时间和频次，如每天、每周或间隔周期等，形成长期定时的可执行任务计划。

### 3.4 自动采集发布接口的可配置化实现

主要实现了三种自动采集发布方式:

(1) 直接发布为本地的 Word、Excel、CSV、MDB 格式文件

支持预先定义发布的格式文件模板和保存位置，然后按模板将文件转化为相应结构格式并保存到指定位置。

(2) 发布到指定的关系数据库表

主要是 MySQL 等主流关系数据库，且关系数据库中表结构必须已存在。通过配置数据库的服务器地址、端口、用户名、密码等登录信息进行自动验证，基于拼接 SQL 语句将已采集的内容标签信息与关系数据库中的表结构进行关联映射，实现采集信息自动提交入库。

(3) 发布到远程 CSpace 知识库系统

前两种方式比较简单，实现关键在于数据格式的转换和元数据字段内容的映射，而第三种方式相对复杂，不仅要能够实现自动向CSpace远程知识库系统实时或定时提交最新采集数据，同时也必须能够实现并保证 CSpace 对已提交采集数据的自动实时正确接收，因此需要着重研究。

CSpace 知识库系统自身具有一套复杂而又完整的数据提交、审核发布、存储索引流程，自动提交到 CSpace 的采集数据必须打通和融入这些流程，才能后续正常应用 CSpace 的数据、权限等各种管理功能，实现与 CSpace 平台自身的各种功能交互以及基于 CSpace 与第三方系统的接口交互等。基于以上考虑，该方法主要分为以下步骤进行实现:

①面向 CSpace 的自动登录验证配置。支持用户在采集发布接口中配置远程 CSpace 知识库系统的登录信息(用户名、密码、验证码等)，接口应用时会自动调用该配置信息和 CSpace 的登录机制，向 CSpace 发出登录请求并进行验证，最后返回登录验证成功与否的标志信息。如 CSpace 4.0 系统登录成功的标志一般是返回信息: {original_url: "/mycspace"}，表示登录成功后转向到“我的工作间”。

②基于数据包方式的已采集数据与 CSpace 元数据的关联映射配置。支持用户将已采集数据的内容标签与 CSpace 元数据字段进行一一对应的映射配置，主要包括采集资源类型与 CSpace 知识对象类型的对应，采集信息内容标签与 CSpace 对象类型模板中元数据字段的一一对应，预定义提交发布的 CSpace 的研究单元/专题名称及 ID 配置等。接口应用时系统会自动调用该映射配置信息，采用 httpclient 提交 post 数据包的方式，将该信息模拟并构造为表单提交数据的方式，向 CSpace 工作流自动提交与确认发布数据。

③基于 CSpace 提交流程改进的数据接收与发布接口类的实现。为便于人工进行数据浏览确认，原有 CSpace 提交流程具有提交缓存和二次确认修改的工作流，相应的数据流会首先存储在 CSpace Workflow 的工作机制中，最终确认提交后才会进入 CSpace 的存储索引流程。在机器自动发布接口实现中，由于不需要提交缓存的工作流并且提交缓存会影响采集发布结果的查重与效率，因此略去了该过程，实现了自动查重与数据处理提交发布的接口类。该接口类需要在 CSpace 的 web.xml 中进行配置，主要包含几个核心方法。其中 processQuickSubmitGather 方法，用于响应提交请求，并进行数据实时接收、任务分发和处理状态返回。processInputFormGather 方法，用于处理接收到的数据，与 CSpace 的元数据字段实现自动关联映射。processKeywordsGather 方法，用于进行数据确认提交前的内容再次过滤，需要传入待过滤的正文内容、关键词列表等参数。可根据需要针对不同的 CSpace 子专题嵌入预定义好的不同的关键词列表或领域词表，通过调用或重写该方法，对已采集到的数据内容进行算法过滤排除，不符合专题要求的数据则不进行提交发布。

### 3.5 应用效果展示

本文结合项目研究需求，首先在海洋科技信息监测中进行实际应用。主要是对海洋科技信息包含如相关海洋国家制定的综合性海洋科技发展战略、海洋政策与法律及相关重要海洋研究机构发布的海洋战略研究报告、资讯报告、统计资料、新闻报道、学术会议等的自动监测配置与采集发布。前期确立的海洋科技情报监测目录涵盖了大约 21 个数据源、35 个站点栏目。用户可在采集任务配置界面新建采集任务模板，配置采集网址规则、内容信息提取规则、数据过滤处理规则等，形成一系列可执行任务，如图 4 所示。同时在自动采集发布配置界面实现了针对 CSpace 的远程自动登录验证配置，与CSpace元数据的关联映射配置等，如图 5 所示。最终经过配置的任务，可长期定时自动监测、采集目标来源站点的最新数据，并实时提交到远程 CSpace 知识库。

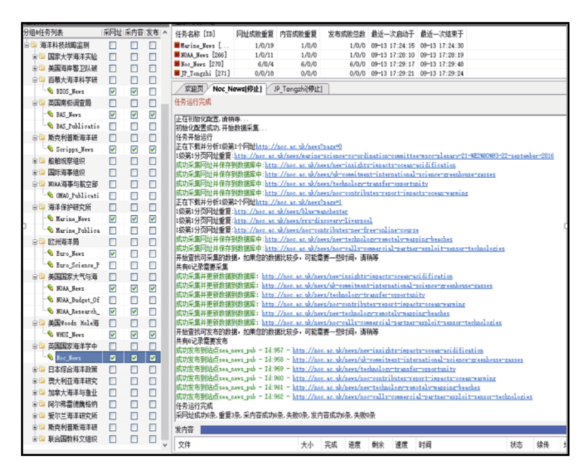


图 4　采集任务运行配置

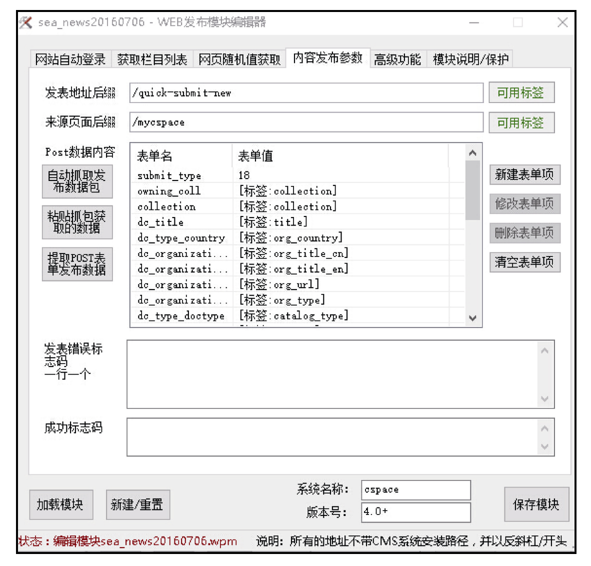



图 5　采集内容发布配置

## 4　结　论

实际应用表明，本文研究较大程度上扩展了 CSpace 的数据采集集成功能，且具有以下优点:

(1) 通用的可配置化的自动监测方法可以应用在多个科技信息监测领域

主要是对具有相同结构内容特质的科技信息资源来源网站如各种门户网站、机构网站、自由网站都可以进行配置以实现自动监测。除此之外，对 OAI-PMH 接口、JSON 接口等各种常规开放接口也能进行监测采集。该自动监测功能已在中国科学院兰州文献情报中心海洋科技战略信息自动监控平台、产业情报大数据平台[23]、全球变化知识资源中心[24]、全球科研项目数据库[25]等多个基于 CSpace 的项目平台建设中得到应用。

(2) 与 CSpace 系统的松耦合性

自动监控功能与 CSpace 通过自动采集发布接口配置实现交互，在整体上形成了一种基于客户端与服务器的 C/S 架构工作模式。自动监控功能作为客户端，可分布式多线程运行在多台机器上进行监测采集以提高采集效率; CSpace 知识库作为远程服务器，用于对已采集数据的实时接收、存储索引与分析管理。当服务端出现问题，至多是采集数据发布不成功，并不影响客户端的数据监测采集，反之亦然。而且，采集任务配置与自动采集发布接口配置可以是多对多的关系，不同的采集任务可以配置不同的采集发布接口，实现同时向多个 CSpace 系统自动采集发布数据。

但同时，本文也存在一定不足:

(1) 采集内容规则配置过程比较复杂: 需要配置人员对 HTML 标签内容、正则表达式、XPath 表达式的构造具有一定的理论知识。主要是针对特定信息源特定栏目的自动监测采集，目标数据源及内容抽取规则需要预先遴选与配置，不支持在自动监测采集过程中自动识别发现新的信息源。同时采集内容规则严格依赖于对科技信息来源站点栏目内容结构的特征分析与模板化创建，当原始内容结构发生变化时，采集内容规则配置也应随之变化。

(2) 暂不支持对一些比较复杂的基于 Ajax 技术或需要登录的来源站点的自动监测: 前期仅实现了对大部分常规开放性科技信息源的监测采集配置，对一些强制需要登录才能获得采集内容的来源站点，仅解决了通过短暂点击获得 Cookie 信息的配置验证，并不具有长久的时效性。

以上问题都需要后续深入研究，期望能够借助机器学习算法、启发式算法完成对自动监测功能的优化，包括简化采集内容规则的配置过程，实现对一些复杂站点的分析与自动采集配置等，从而基于 CSpace 对专题领域科技研究、科技资源采集集成建设等提供更加良好的支持。

**参考文献:**

[1] 祝忠明. 支持数据与知识服务的机构知识库新功能[R/OL]. (2016-10-17). [2017-07-17]. http://ir.las.ac.cn/handle/12502/8879. (Zhu Zhongming. New Functions of Institutional

Repository for Data and Knowledge Services [R/OL]. (2016-10-17). [2017-07-17]. http://ir.las.ac.cn/handle/8879.)

[2] 张晓林. 机构知识库的发展趋势与挑战[J]. 现代图书情报技术, 2014(2): 1-7. (Zhang Xiaolin. Trends and Challenges for Institutional Repositories [J]. New Technology of Library and Information Service, 2014(2): 1-7.)

[3] 姚晓娜, 祝忠明, 刘巍, 等. 机构知识库集成服务系统研究及实践[J]. 图书情报工作, 2015, 59(21): 123-127, 75. (Yao Xiaona, Zhu Zhongming, Liu Wei, et al. Research and Practice on the Institutional Repository Aggregative System [J]. Library and Information Service, 2015, 59(21): 123-127, 75.)

[4] 叶勤勇. 基于URL规则的聚焦爬虫及其应用[D]. 杭州: 浙江大学, 2007. (Ye Qinyong. URL Rule Based Focused Crawl and Its Application[D]. Hangzhou: Zhejiang University, 2007.)

[5] 蒋付彬. 基于决策树的URL分类器算法及主题爬虫平台设计[D]. 成都: 成都理工大学, 2016. (Jiang Fubin. URL Classifier Algorithm Based on Decision Tree and Platform Design of Focused Crawler [D]. Chengdu: Chengdu University of Technology, 2016.)

[6] 杨镒铭. 基于URL模式的网页分类算法研究[D]. 合肥: 中国科学技术大学, 2016. (Yang Yiming. Research on URL-Pattern Based Algorithm for Web Page[D]. Hefei: University of Science and Technology of China, 2016.)

[7] Bar-Yossef Z, Rajagopalan S. Template Detection via Data Mining and Its Applications[C]//Proceedings of the 11th International Conference on World Wide Web, Honolulu, Hawaii, USA. New York, USA: ACM, 2002: 580-591.

[8] Mitra P, Debnath S, Giles Lee C, et al. Automatic Identification of Informative Sections of Web Pages [J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 17(9): 1233-1246.

[9] 王浩. 基于半监督学习的网络敏感信息识别[D]. 天津: 天津大学, 2012. (Wang Hao. Internet Sensitive Information Identification Based on Semi-Supervised Learning [D]. Tianjin: Tianjin University, 2012.)

[10] Pavlinek M, Podgorelec V. Text Classification Method Based on Self-training and LDA Topic Models [J]. Expert Systems with Applications, 2017, 80: 83-93.

[11] 李剑. 基于 DOM 和神经网络的网页净化应用[J]. 电子科技, 2012, 25(1): 105-107. (Li Jian. Application Research of Web Page Purification Based on DOM and Neural Network[J]. Electronic Science and Technology, 2012, 25(1): 105-107.)

[12] 李伟男, 李书琴, 景旭, 等. 基于模拟退火算法和二阶HMM的Web信息抽取[J]. 计算机工程与设计, 2014, 35(4): 1264-1268. (Li Weinan, Li Shuqin, Jing Xu, et al. Web Information Extraction Based on Simulated Annealing Algorithm and Second-order HMM [J]. Computer Engineering and Design, 2014, 35(4): 1264-1268.)

[13] Cai D, Yu S, Wen J R, et al. VIPS: A Vision-based Page Segmentation Algorithm [R]. Microsoft Research, Technical Report MSR-TR-2003-79, 2003.

[14] 谢方立. 基于节点类型标注的网页主题信息提取技术研究[D]. 北京: 中国农业科学院, 2016. (Xie Fangli. Research on the Technique of Extracting Web Page Informational Content Based on Node Type Annotation[D]. Beijing: Chinese Academy of Agricultural Sciences, 2016.)

[15] 欧健文, 董守斌, 蔡斌. 模板化网页主题信息的提取方法[J]. 清华大学学报: 自然科学版, 2005, 45(S1): 1743-1747. (Ou Jianwen, Dong Shoubin, Cai Bin. Topic Information Extraction from Template Web Pages[J]. Journal of Tsinghua University: Science and Technology, 2005, 45(S1): 1743-1747.)

[16] 马费成, 苏小敏. 网络信息生命阶段的模糊识别研究[J]. 情报科学, 2012, 30(9): 1277-1283. (Ma Feicheng, Su Xiaomin. Research on Fuzzy Identification in Life Stages of Network Information [J]. Information Science, 2012, 30(9): 1277-1283.)

[17] 林文辉. 基于 Hadoop 的海量网络数据处理平台的关键技术研究[D]. 北京: 北京邮电大学, 2014. (Lin Wenhui. Research on Key Technologies of Massive Network Data Processing Platform Based on Hadoop [D]. Beijing: Beijing University of Posts and Telecommunications, 2014.)

[18] 谭宗颖, 王强, 苍宏宇, 等. 科技发展前沿信息监测与分析平台的构建[J]. 科学学研究, 2010, 28(2): 195-201. (Tan Zongying, Wang Qiang, Cang Hongyu, et al. Construction of the Science and Technology Frontier Information Monitoring and Analysis Platform [J]. Studies in Science of Science, 2010, 28(2): 195-201.)

[19] 刘海波. 动态 Web 信息监测相关技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2011. (Liu Haibo. Research on Related Technology of Dynamic Web Information Monitoring [D]. Harbin: Harbin Institute of Technology, 2011.)

[20] 张智雄, 刘建华, 谢靖, 等. 科技战略情报监测服务云平台的设计与实现[J]. 现代图书情报技术, 2014(6): 51-61. (Zhang Zhixiong, Liu Jianhua, Xie Jing, et al. Design and Implementation of the Service Cloud for Strategic S&T Information Monitoring [J]. New Technology of Library and Information Service, 2014(6): 51-61.)

[21] 谢靖, 曲云鹏, 刘建华. 面向网络科技监测的分布式定向

资源精确采集研究和应用[J]. 现代图书情报技术, 2011 (7-8): 26-31. (Xie Jing, Qu Yunpeng, Liu Jianhua. Targeted Websites Distributed and Precise Harvest System for Network Monitoring Technology[J]. New Technology of Library and Information Service, 2011(7-8): 26-31.)
[22] 王思丽, 马建玲, 王楠, 等. 开放知识资源的元数据自动采集策略研究[J]. 图书馆学研究, 2013(12): 47-51. (Wang Sili, Ma Jianling, Wang Nan, et al. Research on Automatic Acquisition Strategy for Metadata of Open Knowledge Resources [J]. Research on Library Science, 2013(12): 47-51.)

## 作者贡献声明:

王思丽: 采集内容规则的可配置化实现, 论文起草与修订;
刘巍: 整体功能结构设计, 面向 CSpace 的自动采集发布接口的可配置化实现;
祝忠明: 提出研究思路, 设计研究方案;
吴志强: 负责功能测试与实验研究;
王金平: 调研与确定海洋科技情报源监测目录。

## 利益冲突声明:

所有作者声明不存在利益冲突关系。

## 支撑数据:

支撑数据由作者自存储, E-mail: wangsl@llas.ac.cn。
[1] 王金平, 刘巍, 王思丽. 海洋科技战略信息监测目录数据.xls.

收稿日期: 2017-08-05
收修改稿日期: 2017-08-17

# Tracking Scientific Information with CSpace Technology

Wang Sili[1, 2]　Liu Wei[1]　Zhu Zhongming[1]　Wu Zhiqiang[1]　Wang Jinping[1]
[1](Lanzhou Literature and Information Center, Chinese Academy of Sciences, Lanzhou 730000, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

**Abstract:** [**Objective**] This paper proposes a new system to automatically track, acquire, store and manage scientific information, aiming to support research in related fields. [**Methods**] We developed the new system based on the CSpace and then solve many technical issues. Then, we examined the new system with marine information. [**Results**] The proposed system could automatically retrieve multi-source heterogeneous scientific information, which supported the construction of science and technology platform. [**Limitations**] The information acquisition procedure of the new system was complex, and it cannot retrieve documents from password-protected sites. [**Conclusions**] The proposed method could expand the CSpace's data acquisition and integration functions, and might be transferred to other fields.

**Keywords:** Cspace　Institutional Repository　Scientific and Technological Information　Automatic Monitoring　Information Acquisition